

# Combining p-values via averaging

Vladimir Vovk  
 v.vovk@rhul.ac.uk  
 http://vovk.net

December 27, 2012

## Abstract

This note discusses the problem of multiple testing of a single hypothesis, with a standard goal of combining a number of p-values without making any assumptions about their dependence structure. An old result by Rüschendorf shows that the p-values can be combined by scaling up their average by a factor of 2 (but no smaller factor is sufficient in general).

## 1 Introduction

Suppose we are testing the same hypothesis using  $K \geq 2$  different statistical tests and obtaining p-values  $p_1, \dots, p_K$ . How can we combine them into a single p-value?

One of the earliest papers answering this question was Fisher's [2]. However, Fisher's paper assumes that the p-values are independent, whereas we would like to avoid any assumptions besides all  $p_k$ ,  $k = 1, \dots, K$ , being bona fide p-values. Fisher's method has been extended to dependent p-values in, e.g., [1, 8], but the combined p-values obtained in those papers are approximate; in this note we are interested in precise or conservative p-values.

The simplest method for combining p-values is the Bonferroni method:

$$F(p_1, \dots, p_K) := K \min(p_1, \dots, p_K) \quad (1)$$

(when  $F(p_1, \dots, p_K)$  exceeds 1 it can be replaced by 1, but we usually ignore this trivial step). Albeit  $F(p_1, \dots, p_K)$  is a p-value, it has been argued that in many cases it is overly conservative. Rüger [11] extends the Bonferroni method by showing that, for any fixed  $k \in \{1, \dots, K\}$ ,

$$F(p_1, \dots, p_K) := \frac{K}{k} p_{(k)} \quad (2)$$

is a p-value, where  $p_{(k)}$  is the  $k$ th smallest p-value among  $p_1, \dots, p_K$ ; see [10] for a simpler exposition. Hommel [5] develops this by showing that

$$F(p_1, \dots, p_K) := \left(1 + \frac{1}{2} + \dots + \frac{1}{K}\right) \min_{k=1, \dots, K} \frac{K}{k} p_{(k)} \quad (3)$$

is also a p-value. (Simes [13] has improved (3) by removing the first factor on the right-hand side of (3), but he assumes the independence of  $p_1, \dots, p_K$ .)

Intuitively, the most natural way to combine  $K$  numbers is simply to average them; essentially, this is the way of combining p-values used in the method of cross-conformal prediction (see [14], (11)). None of the functions  $F$  in (1), (2), and (3) involves the average  $\bar{p} := \frac{1}{K}(p_1 + \dots + p_K)$ . This note draws the reader's attention to a result by Rüschendorf ([12], Theorem 1) showing that  $\bar{p}$  is not always a p-value but  $2\bar{p}$  is; moreover, the factor of 2 cannot be improved in general.

Section 2 proves the part of Rüschendorf's result stating that  $2\bar{p}$  is a bona fide p-value (perhaps conservative). Section 3 considers the case  $K = 2$ , in which it is very easy to see that the factor of 2 is optimal.

It is often possible to automatically transform results about multiple testing of a single hypothesis into results about testing multiple hypotheses; the standard procedures are Marcus et al.'s [9] closed testing procedure and its modification by Hommel [6]. In particular, when applied to the Bonferroni method the closed testing procedure gives the well-known method due to Holm [4]; see, e.g., [6, 7] for its further applications. Unfortunately, the closed testing procedure does not appear to lead to a simple and intuitive way of testing multiple hypotheses when combined with Rüschendorf's result, and it will not be discussed further in this note.

## Some notation and terminology

If  $E$  is a property of elements of a set  $X$ ,  $\mathbf{1}_E : X \rightarrow [0, \infty)$  is the indicator function of  $E$ :  $\mathbf{1}_E(x) = 1$  if  $x$  satisfies  $E$  and  $\mathbf{1}_E(x) = 0$  if not. A function  $F : [0, 1] \rightarrow [0, \infty)$  is *increasing* (resp. *decreasing*) if  $F(x_1) \leq F(x_2)$  (resp.  $F(x_1) \geq F(x_2)$ ) whenever  $x_1 \leq x_2$ . A function  $F : [0, 1]^K \rightarrow [0, \infty)$  is *increasing* (resp. *decreasing*) if it is increasing (resp. decreasing) in each of its arguments. A set in  $[0, 1]^K$  is *increasing* (resp. *decreasing*) if its indicator function is increasing (resp. decreasing).

## 2 Combining p-values by scaled averaging

A *p-value function* is a random variable  $P$  that satisfies

$$\mathbb{P}(P \leq \epsilon) \leq \epsilon, \quad \forall \epsilon \in [0, 1].$$

The values taken by a p-value function are *p-values* (allowed to be conservative). (In Section 1 the expression “p-value” was loosely used to refer to p-value functions as well.) A *merging function* is an increasing Borel function  $F : [0, 1]^K \rightarrow [0, \infty)$  such that  $F(U_1, \dots, U_K)$  is a p-value function, where  $U_1, \dots, U_K$  are random variables distributed uniformly on  $[0, 1]$ .

**Remark.** The requirement that a merging function be Borel does not follow automatically from the requirement that it be increasing: see the remark after

Theorem 4.4 in [3] (Theorem 4.4 itself says that every increasing function on  $[0, 1]^K$  is Lebesgue measurable).

Notice that, for any merging function  $F$ ,  $F(P_1, \dots, P_K)$  is a p-value function whenever  $P_1, \dots, P_K$  are p-value functions. Indeed, for each  $k \in \{1, \dots, K\}$  we can define a uniformly distributed random variable  $U_k \leq P_k$  by

$$U_k(\omega) := \mathbb{P}(P_k < P_k(\omega)) + \theta \mathbb{P}(P_k = P_k(\omega)), \quad \omega \in \Omega,$$

where  $\theta$  is a random variable distributed uniformly on  $[0, 1]$ , and  $\Omega$  is the underlying probability space extended (if required) to carry such a  $\theta$ ; we then have

$$\mathbb{P}(F(P_1, \dots, P_K) \leq \epsilon) \leq \mathbb{P}(F(U_1, \dots, U_K) \leq \epsilon) \leq \epsilon, \quad \forall \epsilon \in [0, 1].$$

The following proposition states Rüschendorf's result in terms of merging functions.

**Proposition 1.** *The function  $M : [0, 1]^K \rightarrow [0, 1]$  defined by*

$$M(p_1, \dots, p_K) := \frac{2}{K}(p_1 + \dots + p_K) \tag{4}$$

*is a merging function.*

The rest of this section is devoted to a self-contained proof of Proposition 1. A *copular probability measure* is a probability measure on  $[0, 1]^K$  all of whose marginals are uniform probability measures on  $[0, 1]$ . The *upper copular probability*  $\mathbb{C}(E)$  of a Borel set  $E \subseteq [0, 1]^K$  is defined to be the supremum of  $x(E)$ ,  $x$  ranging over the copular probability measures. In terms of  $\mathbb{C}$ , an increasing Borel function  $F : [0, 1]^K \rightarrow [0, \infty)$  is a merging function if and only if  $\mathbb{C}(F \leq \epsilon) \leq \epsilon$  for all  $\epsilon \in [0, 1]$ . We say that a merging function  $F$  is *precise* if  $\mathbb{C}(F \leq \epsilon) = \epsilon$  for all  $\epsilon \in [0, 1]$ .

For  $s \in [0, \infty)$ , define

$$E_s := \{(u_1, \dots, u_K) \in [0, 1]^K \mid u_1 + \dots + u_K \leq s\} \subseteq [0, 1]^K. \tag{5}$$

Proposition 1 can be strengthened: in fact,  $M$  is a precise merging function. The original statement of this result is as follows.

**Lemma 1** ([12], Theorem 1). *For any  $s \in [0, \infty)$ ,*

$$\mathbb{C}(E_s) = \min\left(\frac{2s}{K}, 1\right).$$

**Remark.** In Section 1 we already alluded to an example of a set with a known upper copular probability: the set

$$\{(u_1, \dots, u_K) \in [0, 1]^K \mid \mathbf{1}_{u_1 \leq \alpha} + \dots + \mathbf{1}_{u_K \leq \alpha} \geq k\},$$

where  $\alpha \in [0, 1]$  and  $k \in \{1, \dots, K\}$ , has upper copular probability of  $(K/k)\alpha$ ; this is equivalent to (2) being a merging function. Another well-known example is  $H := [0, u_1] \times \dots \times [0, u_K]$ , where  $u_1, \dots, u_K \in [0, 1]$ . The upper copular probability of  $H$  is  $\min(u_1, \dots, u_K)$ . This is known as one of the Fréchet–Hoeffding bounds in the theory of copulas. Lemma 1 is one more example of this kind. Lemma 2 below will give a simple characterization of upper copular probability in the easy case  $K = 2$ .

Given Lemma 1, the proof of Proposition 1 is trivial: for any  $\epsilon \in [0, 1]$ ,

$$\mathbb{P}(M(U_1, \dots, U_K) \leq \epsilon) = \mathbb{P}\left(U_1 + \dots + U_K \leq \frac{K\epsilon}{2}\right) \leq \epsilon.$$

Notice that for the proof of Proposition 1 we only need the inequality  $\leq$  in the lemma. The rest of this section is devoted to the proof of this inequality.

Let  $K[0, 1]$  be the sum of  $K$  disjoint copies of the interval  $[0, 1]$ . A (somewhat arbitrary) concrete representation of  $K[0, 1]$  is the set  $\cup_{k=1}^K [2(k-1), 2k-1]$ . We will sometimes use the notation  $K[0, 1]_k$  for the  $k$ th copy of  $[0, 1]$  in  $K[0, 1]$ ; so that  $K[0, 1]_k = [2(k-1), 2k-1]$  in the concrete representation (but  $K[0, 1]_k$  is always identified with  $[0, 1]$ , via the bijection  $u \mapsto u - 2(k-1)$  in the concrete representation). If  $x$  is a measure on  $[0, 1]^K$ , we define  $x_k$  to be the projection of  $x$  onto the  $k$ th coordinate of  $[0, 1]^K$ ,

$$x_k(E) := x([0, 1]^{k-1} \times E \times [0, 1]^{K-k}), \quad E \subseteq [0, 1] \text{ is Borel},$$

and we define  $Ax$  to be the measure on  $K[0, 1]$  that coincides with  $x_k$  on  $K[0, 1]_k$  (so that  $Ax$ 's total mass is  $K$  when  $x$  is a probability measure). The *uniform measure* on  $K[0, 1]$  is the measure on the Borel  $\sigma$ -algebra on  $K[0, 1]$  that coincides with the uniform probability measure on each of its components  $K[0, 1]_k$  (so that  $Ax$  is the uniform measure on  $K[0, 1]$  if and only if  $x$  is a copular probability measure).

Lemma 1 can be interpreted as a statement about the following infinite-dimensional problem of linear programming:

$$cx \rightarrow \sup \quad \text{subject to} \quad Ax = b, \quad x \geq 0, \quad (6)$$

where  $c$  is the indicator function of the set  $E_s$ , the variable  $x$  ranges over all measures on  $[0, 1]^K$ ,  $cx$  is understood to be  $\int c dx$ ,  $Ax$  is as defined above, and  $b$  is the uniform measure on  $K[0, 1]$ . The condition  $x \geq 0$  is an embellishment without a formal meaning (and emphasizes the fact that measures take only nonnegative values). Lemma 1 says that the value of (6) is  $2s/K$ .

The formal dual problem to (6) is

$$\lambda b \rightarrow \inf \quad \text{subject to} \quad \lambda A \geq c, \quad (7)$$

which we will interpret as follows: the dual variable  $\lambda$  ranges over all Borel functions on  $K[0, 1]$ ,  $\lambda b$  is understood to be  $\int \lambda db$ ,  $\lambda A$  is the function on  $[0, 1]^K$  defined by

$$(\lambda A)(u_1, \dots, u_K) := \lambda_1(u_1) + \dots + \lambda_K(u_K),$$

where  $\lambda_k$  is the restriction of  $\lambda$  to  $K[0, 1]_k$ , and  $\geq$  stands, as usual, for the pointwise inequality.

It is easy to see that the operators  $x \mapsto Ax$  and  $\lambda \mapsto \lambda A$  are dual, in the sense that  $(\lambda A)x = \lambda(Ax)$ :

$$(\lambda A)x = \int \lambda_1 dx_1 + \cdots + \int \lambda_K dx_K = \int \lambda dAx = \lambda(Ax).$$

(This justifies using the same letter for both operators.) As usual, the value of the original problem (6) does not exceed the value of the dual problem (7): indeed, if  $x$  satisfies the constraints in (6) and  $\lambda$  satisfies the constraint in (7),

$$cx \leq (\lambda A)x = \lambda(Ax) = \lambda b.$$

Now we have all components for the proof of the inequality  $\leq$  in Lemma 1.

*Proof of the inequality  $\leq$  in Lemma 1.* It suffices to prove that the value of the dual problem (7) does not exceed  $2s/K$ . Define  $\lambda : K[0, 1] \rightarrow [0, \infty)$  by  $\lambda_k(u) := (2/K - u/s)^+$  for all  $k \in \{1, \dots, K\}$ , where  $t^+$  is  $t$  if  $t \geq 0$  and 0 otherwise. (In other words, assuming  $s \leq K/2$ ,  $\lambda_k : [0, 1] \rightarrow [0, \infty)$  is the function with the subgraph of the smallest area among all functions that are linear when positive and whose graph passes through  $(s/K, 1/K)$ .) Since

$$\lambda b = \int \lambda_1 + \cdots + \int \lambda_K \leq 2s/K$$

(with  $=$  in place of the last  $\leq$  when  $s \leq K/2$ ), it remains to prove that the constraint in (7) is satisfied. This is accomplished by the following chain of inequalities:

$$\begin{aligned} \lambda A(u_1, \dots, u_K) &= \sum_{k=1}^K \left( \frac{2}{K} - \frac{u_k}{s} \right)^+ \geq \left( \sum_{k=1}^K \left( \frac{2}{K} - \frac{u_k}{s} \right) \right)^+ \\ &= (2 - (u_1 + \cdots + u_K)/s)^+ \geq \mathbf{1}_{2 - (u_1 + \cdots + u_K)/s \geq 1} = \mathbf{1}_{u_1 + \cdots + u_K \leq s} = c(u_1, \dots, u_K). \end{aligned}$$

□

### 3 Case $K = 2$

In the case  $K = 2$  upper copular probability admits a simple characterization.

**Lemma 2.** *If a nonempty Borel set  $E \subseteq [0, 1]^2$  is decreasing, its upper copular probability is*

$$\mathbb{C}(E) = \min \left( \inf \{ u_1 + u_2 \mid (u_1, u_2) \in [0, 1]^2 \setminus E \}, 1 \right). \quad (8)$$

Lemma 2 implies that the factor 2 in (4) is optimal for  $K = 2$ : indeed, it shows that the function  $M_\alpha(p_1, p_2) := \alpha(p_1 + p_2)$ , where  $\alpha > 0$ , satisfies  $\mathbb{C}(M_\alpha \leq \epsilon) = \min(\epsilon/\alpha, 1)$  for all  $\epsilon \in [0, 1]$ ; therefore,  $M_\alpha$  is a merging function if and only if  $\alpha \geq 1$ . It is clear that  $M_1 = M$  is the only precise merging function among  $M_\alpha$ .

*Proof of Lemma 2.* Let  $E$  be a nonempty decreasing Borel set in  $[0, 1]^2$ ; suppose  $\mathbb{C}(E)$  is strictly less than the right-hand side of (8). Let  $t$  be any number strictly between  $\mathbb{C}(E)$  and the right-hand side of (8). The copular probability measure concentrated on

$$[(t, 0), (0, t)] \cup [(t, 1), (1, t)]$$

has a value of at least  $t$  on  $E$  since  $E$  contains  $[(t, 0), (0, t)]$ . Therefore,  $\mathbb{C}(E) \geq t$ . This contradiction proves the inequality  $\geq$  in (8).

The inequality  $\leq$  in (8) follows from Lemma 1 (part  $\leq$ ). Indeed, denoting the right-hand side of (8) as  $s$  and assuming  $s < 1$  (the case  $s = 1$  is trivial), we have  $E \subseteq E_{s+\epsilon}$  for an arbitrarily small  $\epsilon > 0$ , in the notation of (5). Therefore, by Lemma 1,  $\mathbb{C}(E) \leq \mathbb{C}(E_{s+\epsilon}) \leq s + \epsilon$ .  $\square$

A merging function  $F_1$  *dominates* a merging function  $F_2$  if  $F_1 \leq F_2$ . The following corollary of Lemma 2 says that, in the case  $K = 2$ , the merging function (4) is dominated by all precise merging functions. This is not true when  $K > 2$ : for example, for the Bonferroni function (1) we have  $M(p, \dots, p) = 2p < Kp = F(p, \dots, p)$ .

**Corollary 1.** *When  $K = 2$ , any precise merging function dominates  $M$ .*

*Proof.* Let  $F : [0, 1]^2 \rightarrow [0, \infty)$  be a merging function that does not dominate  $M$ . Choose  $(u_1, u_2) \in [0, 1]^2$  such that  $F(u_1, u_2) > u_1 + u_2$  and choose  $\epsilon \in (u_1 + u_2, F(u_1, u_2))$ . Since  $\{F \leq \epsilon\}$  does not contain  $(u_1, u_2)$ , we have  $\mathbb{C}(F \leq \epsilon) \leq u_1 + u_2 < \epsilon$ , and so  $F$  is not precise.  $\square$

## Acknowledgments

I am grateful to Dave Cohen and Alessio Sancetta for their advice. This work was partially supported by the Cyprus Research Promotion Foundation.

## References

- [1] Morton B. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31:987–992, 1975.
- [2] Ronald A. Fisher. Combining independent tests of significance. *American Statistician*, 2:30, 1948.
- [3] B. T. Graham and G. R. Grimmett. Influence and sharp-threshold theorems for monotonic measures. *Annals of Probability*, 34:1726–1745, 2006.
- [4] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.
- [5] G. Hommel. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, 25:423–430, 1983.

- [6] G. Hommel. Multiple test procedures for arbitrary dependence structures. *Metrika*, 33:321–336, 1986.
- [7] G. Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383–386, 1988.
- [8] James T. Kost and Michael P. McDermott. Combining dependent p-values. *Statistics and Probability Letters*, 60:183–190, 2002.
- [9] R. Marcus, E. Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660, 1976.
- [10] D. Morgenstern. Berechnung des maximalen Signifikanzniveaus des Testes “Lehne  $H_0$  ab, wenn  $k$  unter  $n$  gegebenen Tests zur Ablehnung führen”. *Metrika*, 27:285–286, 1980.
- [11] B. Rüger. Das maximale Signifikanzniveau des Testes “Lehne  $H_0$  ab, wenn  $k$  unter  $n$  gegebenen Tests zur Ablehnung führen”. *Metrika*, 25:171–178, 1978.
- [12] Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14:623–632, 1982.
- [13] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.
- [14] Vladimir Vovk. Cross-conformal predictors. Technical Report [arXiv:1208.0806v1](https://arxiv.org/abs/1208.0806v1) [stat.ML], [arXiv.org](https://arxiv.org/) e-Print archive, August 2012.